Using a hierarchical segmented model to assess the dynamics of leaf appearance in plant populations.

Charlotte Baey, Paul-Henry Cournède

Using a hierarchical segmented model to assess the dynamics of leaf appearance in plant populations

Charlotte Baey[1 2] and Paul-Henry Cournède[1 2]

[1] Laboratoire MAS, Ecole Centrale Paris
[2] EPI Digiplante, INRIA Saclay Ile-de-France

Abstract Modeling inter-individual variability in plant populations is a key issue to enhance the predictive capacity of plant growth models at field level. In sugar beet, this variability is well illustrated by the phyllochron (thermal time elapsing between two successive leaf appearances): even if the mean phyllochron remains stable within a given variety, there is a high heterogeneity between individuals. When considering the dynamics of leaf appearance as a function of thermal time in sugar beet, two linear phases can be observed, leading to the definition of a hierarchical segmented model with four random parameters varying from one individual to another: thermal time of initiation, first phyllochron, rupture thermal time and second phyllochron. The SAEM-MCMC algorithm is used to estimate the model parameters.
Keywords: nonlinear mixed model, segmented regression, variability, sugar beet.

1 Introduction

A new trend in crop modeling is the development of individual-based plant growth models such as functional-structural plant models (FSPM), combining the description of plant architecture and its physiological functioning (Vos *et al.*[22], de Reffye *et al.*[5]).

The extrapolation of individual-based models to the field scale is still at its early stages. Due to the complexity to describe the growth of all individual plants in the field, an average plant is usually considered for the model calibration process and prediction at field level, as illustrated by Lemaire *et al.*[12] for sugar beet. Beside the difficulty of assessing this 'average plant', the strong variability among individuals makes the average plant prediction quite restrictive, since it only gives a partial characterization of field production.

This variability could arise from three different sources identified by de Reffye *et al.*[6]: the growth initial conditions (typically, the seed weight and the thermal time of initiation), the environmental conditions (the growth of a given plant varies with local environmental conditions), and the genetic variability (in the case of sugar beet, populations are not pure bred, resulting in genetic variations between individuals).

In this paper, we focus on the phenomenon of plant organogenesis (*i.e.* the dynamics of appearance of plant organs) since its right description is crucial for the development of FSPM and since it is affected by several sources of variability. For most crops, the thermal time (corresponding to the accumulation of daily temperature above a base temperature) elapsing between the successive appearances of two phytomers (the elementary structural unit of a plant) usually shows a good range of stability, even though it may be strongly affected by environmental factors, see Clerget *et al.*[2] for a short review on the main results for different plants. In given conditions, the phyllochron may reveal constant for each individual of a population, while showing a strong variability between individuals within a cultivar, see Frank and Bauer[8]. This is typically the case for sugar beet (see Lemaire *et al.*[12]), which is not pure-bred. Likewise, seedling emergence may strongly vary within a population, potentially inducing important variations in the final yield, see Liu *et al.*[15], if the younger plants do not manage to close the gap with the older ones. The development of late plants could also be slowed down by the shadow resulting from the extra-leaves of the early plants.

The sugar beet plant is specifically chosen as test plant for our study. As shown by Lemaire *et al.*[12], and already observed by Milford *et al.*[17], two phases can be observed in the development of new leaves by the sugar beet, leading to the definition of two different phyllochrons. We will thus study the inter-individual variability of the thermal time of initiation (seedling emergence), of the two phyllochrons (controlling the rate of leaf appearance in these two phases), and of the rupture thermal time, i.e. the transition thermal time corresponding to the setting up of the second phase.

To take into account these two phases, a segmented regression will be used, with an unknown break-point corresponding to the rupture thermal time. The variability of the parameters will be assessed using a nonlinear mixed model. We first introduce our methodology of analysis. In section 2, a description of the hierarchical segmented regression model is given, and inference method is described in section 3. We then apply the methodology to experimental data for the sugar beet in section 4, and we also show how the model can be used to build comparisons between different plant populations.

2 The model

Nonlinear mixed models are of particular interest for the analysis of repeated measures data, in many research fields such as pharmacokinetics, agriculture, epidemiology ... In such models, the functional form of the model linking the response variable to time (in the case of longitudinal data) is the same for all individuals, but some parameters are allowed to vary among individuals (see Lindstrom and Bates[14], Davidian and Giltinan[3], Pinheiro and Bates[21] and the references therein).

The sugar beet organogenesis model can be described as a two-stage hierarchical model. In the first stage, the number of leaves according to the thermal time is modeled for a given plant. In the second stage, the variability between plants is assessed by considering each parameter as a random variable.

First stage : intra-individual variation

Let y_{ij} denote the number of leaves of plant i at thermal time t_j. Then we have:

$$y_{ij} = f(t_j, \phi_i) + e_{ij} \tag{1}$$

with ϕ_i the vector of parameters specific to individual i, and e_{ij} a random error term following a normal distribution $\mathcal{N}(0, \sigma^2)$. Thus, f characterize the systematic variation whereas e_{ij} represents the random variation of measurements from individual i.

In our case, f is a two-linear phases function defined as follow:

$$f(t_j, \phi_i) = \phi_{1,i}(t_j - \phi_{0,i}) + \phi_{3,i}(t_j - \phi_{2,i})\ \mathbb{1}_{t_j \geq \phi_{2,i}} \tag{2}$$

with $\phi_{0,j}$ the thermal time of initiation, $\phi_{1,i}$ the first slope (the inverse of the first phyllochron), $\phi_{2,i}$ the rupture thermal time and $\phi_{3,i}$ the difference in slopes between the two phases for plant i. The parameter vector is $\phi_i = (\phi_{0,i}, \phi_{1,i}, \phi_{2,i}, \phi_{3,i})$.

With this formulation, we model the change in slopes between the two phases, rather than the two distinct slopes, and we force the two lines to join at the rupture thermal time.

Second stage : inter-individual variation

In this stage, the inter-individual variation is taken into account by the following model for the parameter vector ϕ_i :

$$\phi_i = \beta + b_i, \quad b_i \sim \mathcal{N}(0, D) \tag{3}$$

where $\beta = (\beta_0, \beta_1, \beta_2, \beta_3)$ is a vector of fixed parameters, b_i is a vector of random effects associated with individual i, and D is a 4×4 covariance matrix. It is assumed to be diagonal with elements ω_0^2, ω_1^2, ω_2^2 and ω_3^2.

This formulation corresponds to the one described by Lindstrom and Bates[14], although more general forms can be used to model the inter-individual variations (see Davidian and Giltinan[3]). Similarly, we make here the normal assumption for the random components b_i, but a more relaxed hypothesis can be assumed.

3 Inference method

A lot of inference methods have been proposed for the nonlinear mixed models, most of them based on maximum likelihood estimation, with a likelihood function based on the joint density of the observations given the covariates.

However, because of the nonlinearity of f, this density has an integral form which is in general analytically intractable.

A list of the different methods can be found in a review by Davidian and Giltinan[4]. The most popular methods are based on an approximation of the likelihood, and are well implemented in classical statistical softwares (for example via the SAS procedure `proc nlmixed` or the S-PLUS/R package `nlme`). However, the likelihood approximation on which they are based may be poor, for example if the number of observations per subject n_i is not large enough or when the Gaussian assumption no longer holds (see Makowski and Lavielle[16]).

An alternative to these methods is to use "exact" or "direct" methods, in which the likelihood function is maximized directly using EM-algorithm, for example (see Walker[23]). Advances in computational power have made these methods more and more appealing. Delyon *et al.*[7] proposed a stochastic approximation of the EM-algorithm, which converges under very general conditions to a local maximum of the function. Kuhn and Lavielle[10] showed that the convergence to maximum likelihood estimates holds when the algorithm is coupled to a MCMC procedure. This method has the advantage of being quicker than an usual EM-algorithm (see Kuhn and Lavielle[11]), so that the convergence can be obtained within a few seconds, and observed likelihood and Fisher Information Matrix could also be estimated. The SAEM-MCMC is implemented in the free software MONOLIX[19].

4 Results

In 2008, field experiments were conducted in France, under three different density conditions: low (5.4 plants per m^2), standard (10.9 plants per m^2) and high (16.4 plants per m^2). The number of leaves of the 45 sugar beet crops (15 plants in each density) was collected at a maximum of 14 different dates. Daily mean values of air temperature (°C) were obtained from French meteorological advisory services (Météo France) five kilometers away from the experimental site to compute the thermal time after sowing.

The model was first applied to the reference density in sugar beet crops, and then a comparison between the three densities was performed. It has been shown by Milford *et al.*[17] and observed by Lemaire *et al.*[12] that the phyllochron of the first phase of development remains quite constant among plant densities, whereas the duration of this first phase was subject to change. This change in phyllochron is observed approximately at full leaf cover, when the competition for light increase. In our model, we thus assumed that the thermal time of initiation and the first slope do not depend on the population density, but we let the rupture thermal time and the second phyllochron vary among the cropping densities.

To estimate the values of $\theta = (\beta, \sigma, D)$, a set of initial values is required for β. We used a first set of values from a previous study of leaf appearance in

sugar beet, described in Lemaire[13]. Other sets of initial values were tested, leading to similar results. Results of the model with the highest likelihood are presented in Table 1.

Parameter	Estimate	Standard error
β_0	99	13
β_1	0.0277	0.00069
β_2	894	79
β_3	-0.0131	0.0013
ω_0	7.58	9.1
ω_1	0.00077	0.00037
ω_2	281	67
ω_3	0.00422	0.0012
σ	1.45	0.09
Log-likelihood	-401.10	

Table1. Results of the SAEM algorithm for the standard density (10.9 pl/m^2)

The average thermal time of initiation in the population is estimated at 99°Cdays, and the average rupture thermal time is estimated at 894°Cdays. The two mean phyllochrons can be easily calculated by taking the inverse of the slopes, and their variances can be approximated using the delta method: if X is a random variable

$$\mathrm{Var}\left(\frac{1}{X}\right) \approx \frac{\mathrm{Var}(X)}{E(X)^4} \tag{4}$$

We find for the first phyllochron a mean value of 36.1°Cdays and a standard deviation of 1°Cdays in the population, and for the second phyllochron, a mean value of 68.5°Cdays and a standard deviation of 6.9°Cdays in the population (we calculated the variance of the second slope by taking the sum of the two variances ω_1^2 and ω_3^2, assuming that the two parameters ϕ_1 and ϕ_3 are independent).

To compare the different densities, two dummy variables $I_{s,i}$ and $I_{h,i}$ were added to the model, equal to 1 if the i-th plant grew in standard and high density respectively, and 0 otherwise.

The function f remains unchanged, but the parameters model is now:

$$\phi_i = A_i\beta + b_i, \quad b_i \sim \mathcal{N}(0, D) \tag{5}$$

with:

$$A_i = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & I_{s,i} & I_{h,i} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & I_{s,i} & I_{h,i} \end{pmatrix} \tag{6}$$

and $\beta = (\beta_0, \beta_1, \beta_{2,l}, \delta_{2,s}, \delta_{2,h}, \beta_{3,l}, \delta_{3,s}, \delta_{3,h})^t$.

Parameters $\delta_{2,s}, \delta_{3,s}$, (resp. $\delta_{2,h}, \delta_{3,h}$) represent the difference between the parameter values in the standard density (resp. the high density) and the parameter values in the low density condition. Wald tests can be used to test for a significant difference between the three densities. Different sets of initial values were used for the parameter vector β, and results of the model with the highest likelihood are presented in Table 2.

Parameter	Estimate	Standard error	Wald tests (p-value)
β_0	70.3	16	-
β_1	0.0267	0.00064	-
$\beta_{2,l}$	1280	7.6	-
$\delta_{2,s}$	-279	22	< 0.0001
$\delta_{2,h}$	-494	15	< 0.0001
$\beta_{3,l}$	-0.0132	0.0012	-
$\delta_{3,s}$	-0.00289	0.0013	0.022
$\delta_{3,h}$	-0.00182	0.0012	0.13
ω_0	13.6	6	-
ω_1	0.00246	0.0003	-
ω_2	8.11	5.3	-
ω_3	0.00234	0.00044	-
σ	1.53	0.052	-
Log-likelihood		-1167.9	

Table2. Comparison of the three densities.

No differences were observed between the standard and the high density. The rupture thermal time was significantly different between low and standard density (1280°Cdays vs. 1001°Cdays), and between low and high density (1280°Cdays vs. 786°Cdays). The second phyllochron was estimated at 74.1°Cdays for the low density, at 61°Cdays for the standard density, and at 65.3°Cdays for the high density. A significant difference for the second slope between standard and low density was observed, but not between standard and high density.

5 Conclusion

The hierarchical segmented regression model used here allows for a better handling of the plant heterogeneity, and thus a better statistical description of the population. All the available data can be used as we no longer resort to an average plant, and comparison between different cropping densities can be performed, taking into account the variability between plants. The mean population value of the parameters and their inter-individual variability can be used as inputs of functional-structural plant models, as described in de Reffye *et al.*[6]. The main issue is then to compute the propagation of these sources of probabilistic uncertainty in dynamic systems of plant growth. This

issue has already been studied in the context of crop models by Monod *et al.*[18]. However, the proposed method to study the output distribution is based on Monte Carlo sampling. It bears strong limitations for assessing individual variability by inverse methods, since it would be necessary to run the Monte Carlo simulations a large number of times. There exist several other methods widely used in other domains (like unscented filtering, see Julier *et al.*[9] and Bevington and Robinson[1]) that overcome this difficulty. The next step of our work is the implementation of these methods to fully develop a full functional-structural plant population model and the associated estimation procedures.

6 Acknowledgement

We are grateful to ITB (sugar beet research institute, Paris) for providing the experimental data to test our methodology. This study is funded by the research project CANTIA (competitiveness cluster IAR, Picardie and Champagne-Ardennes regions).

References

1.P.R. Bevington and D.K. Robinson. *Data reduction and error analysis for the physical science*, McGraw-Hill, Boston, 2003.
2.B. Clerget, M. Dingkuhn, E. Goz, H.F. Rattunde and B. Ney. Variability of Phyllochron, Plastochron and Rate of Increase in Height in Photoperiod-sensitive Sorghum Varieties. *Annals of Botany*, 101, 4, 579–594, 2008.
3.M. Davidian and D. Giltinan. *Nonlinear Models for Repeated Measurement Data*, Chapman and Hall, New-York, 1995.
4.M. Davidian and D. Giltinan. Nonlinear Models for Repeated Measurment Data: An Overview and Update. *Journal of Agricultural, Biological, and Environmental Statistics*, 8, 4, 387–419, 2003.
5.P. de Reffye, P. Heuvelink, E. Barthlmy and P.H. Cournède. Plant Growth Models. *Ecological Models. Vol. 4 of Encyclopedia of Ecology (5 volumes)*, Elsevier, Oxford, 2008.
6.P. de Reffye, S. Lemaire, N. Srivastava, F. Maupas and P.H. Cournède. Modeling Inter-individual Variability in Sugar Beet Populations. *Plant Growth Modeling*, 9, 270–276, 2009.
7.B. Delyon, M. Lavielle and E. Moulines. Convergence of a stochastic approximation version of the EM algorithm. *Annals of Statistics*, 27, 1, 94–128, 1999.
8.A.B. Frank and A. Bauer. Phyllochron Differences in Wheat, Barley, and Forage Grasses. *Crop Science*, 35, 19–23, 1995.
9.S. Julier, J. Uhlmann and H.F. Durrant-Whyte. A new method for the nonlinear transformation of means and covariances in filters and estimators. *IEEE Transactions on Automatic Control*, 45, 3, 477–482, 2000.
10.E. Kuhn and M. Lavielle. Coupling a stochastic approximation version of EM with an MCMC procedure. *ESAIM: Probability and Statistics*, 8, 115–131, 2004.

11.E. Kuhn and M. Lavielle. Maximum likelihood estimation in nonlinear mixed effects models. *Computational Statistics and Data Analysis*, 49, 4, 1020–1038, 2005.
12.S. Lemaire, F. Maupas, P.H. Cournède and P. de Reffye. A morphogenetic crop model for sugar beet (Beta vulgaris L.). *International Symposium on Crop Modeling and Decision Support: ISCMDS*, 5, 19–22, 2008.
13.S. Lemaire. *Systme dynamique de la croissance et du dveloppement de la betterave sucrire (Beta vulgaris L.)*, PhD Thesis, AgroParisTech, 2010.
14.M.J. Lindstrom and D.M. Bates. Nonlinear Mixed Effects Models. *Biometrics*, 46, 673–687, 1990.
15.W. Liu, M. Tollenaar, G. Stewart and W. Deen. Response of Corn Grain Yield to Spatial and Temporal Variability in Emergence. *Crop Science*, 44, 19–23, 2004.
16.D. Makowski and M. Lavielle. Using SAEM to estimate parameters of models of response to applied fertilizer. *Journal of Agricultural, Biological, and Environmental Statistics*, 11, 1, 45–60, 2006.
17.G.F.J. Milford, T.O. Pocock and J. Riley. An analysis of leaf growth in sugar beet. I. Leaf appearance and expansion in relation to temperature under controlled conditions. *Annals of Applied Biology*, 106, 163–172, 1985.
18.H. Monod, C. Naud and D. Makowski. Uncertainty and Sensitivity Analysis for Crop Models. *Working with Dynamic Crop models*, Elsevier, 55–100, 2006.
19.Monolix Team. *User Guide to Monolix*. 2010
20.V.M.R. Muggeo. Estimating regression models with unknown break-points. *Statistics in Medecine*, 22, 19, 3055–3071, 2003.
21.J.C. Pinheiro and D.M. Bates. Approximations to the Log-Likelihood Function in the Nonlinear Mixed-Effects Model. *Journal of Computational and Graphical Statistics*, 4, 1, 12–35, 1995.
22.J. Vos, L.F.M. Marcelis, P. de Vissers, P. Struik and J.B. Evers. *Functional-Structural plant modeling in crop production.* Springer, Berlin, 2007.
23.S. Walker. An EM Algorithm for Nonlinear Random Effects Models. *Biometrics*, 52, 3, 934–944, 1996.